SSPress

Journal of Globe Scientific Reports

journal homepage: www.journal-gsr.online

Paper Type: Original Article

High-Resolution Fire Image Generation via SCGAN-Controlled Methods and Optimized DDPM Models

Lanyan Yang[1], **Yuanhang Cheng**[1,*], **Fang Xu**[2,3], **Boning Li**[2,3,4] , **Xiaoxu Li**[2,3,4]

[1] School of Information Engineering, Shenyang University, Liaoning Shenyang 110044, China
[2] Shenyang Fire Research Institute of M.E.M., Shenyang Liaoning 110034, China
[3] National Engineering Research Center of Fire and Emergency Rescue., Shenyang Liaoning 110034, China
[4] Key Laboratory of Fire Prevention Technology, Liaoning Provincial., Shenyang Liaoning 110034, China
*Corresponding author: Yuanhang Cheng

Abstract

Existing deep learning-based image fire detection algorithms require training the model with a large number of diverse image datasets and accurate annotations to achieve high accuracy and strong anti-interference capabilities. However, in the field of fire detection, there is a lack of sufficiently rich datasets of real fire scenes to train the detection models, leading to unreliable detection results. In this paper, we propose a new flame image generation method that aims to enhance the efficiency and adaptability of fire detection systems, particularly when the number of samples is unbalanced. By constructing extensive datasets containing different environments (e.g., factories, warehouses, and forests), we address the practical challenges of safety control and fire initiation.

Our approach is based on two main networks: the flame generation network and the hybrid network. The flame generation network utilizes the SCGAN technique to generate diverse flame images by controlling the shape of flames based on the input reference information. The hybrid network synthesizes fire images from different scenes into an improved DDPM to create realistic images by fine-tuning textures and styles. Our approach has three main advantages: the ability to control the generated flame images, the preservation of high-quality background details, and training on real datasets, making the generated images suitable for engineering application scenarios.

Keywords: Generative adversarial network,, DDPM Models, Image blending, Image synthesis

1. Introduction

Generating fire imagery is an essential technique for mitigating the disparity in sample counts between images depicting fires and those that do not, a common issue in numerous deep learning applications. To bolster the efficacy and adaptability of fire detection systems, it is imperative to compile extensive datasets from diverse environments including factories, warehouses, and forests. Despite the practical hurdles of safely controlling and initiating fires, the expansion of the fire image database through simulated images is a pressing issue that needs to be addressed.

Existing deep learning-based image fire detection algorithms need to train the model with a large number of rich image datasets and accurate annotations in order to obtain a target detection model with high accuracy and anti-interference capability.

The paper presents an innovative approach to producing flames by manipulating the properties of fire. It draws inspiration from several existing studies and is meant to be flexible enough to function in different settings. The model suggested in this study consists of two networks. The primary contributions are summarized as follows:

1.A flame generation network is proposed to generate a variety of flames using SCGAN and to synthesize fire pictures of various settings by regulating the shape of the result based on any input provided as a reference.

2.In order to generate realistic images by fine-tuning the texture and style of the hybrid region, a hybrid network is proposed to synthesize fire pictures from various scenarios into an enhanced DDPM. It is capable of producing fire samples for engineering application scenarios that can be utilized by target detection algorithms.

The proposed method has three main advantages. First, the fire in the generated fire kernel can be controlled by SCGAN to generate various fire images that cannot be done by existing methods, and to realize the style conversion between fire-free and fire scenes. Second, the flame and background images are further fused by the improved DDPM network, and high quality background details are preserved in the generated images. Finally, since the model is trained on a real dataset, the generated images can be used to generate fire samples for engineering application scenarios that can be used by target detection algorithms.

2. Related Work

Generative image modeling has a rich history in the field of computer vision[31, 10, 15] and has undergone significant advancements in numerous directions during the deep learning era[11].Variational auto-encoders[17], generative adversarial neural networks (GAN) [22], and diffusion models[25] are three significant developments that have substantially improved the learning and modeling capabilities in the deep learning era.

Generative adversarial learning has experienced exponential growth[21, 22, 3, 16, 9], despite the fact that numerous GAN models continue to be challenging to train. VAE models[17] are simpler to train; however, the resulting image quality is frequently illegible. In recent years, diffusion generative models [25, 13, 26, 7] have become increasingly popular due to the exceptional quality of the images they produce[23]. Although generative diffusion models possess exceptional modeling capabilities, they continue to encounter obstacles in both training and synthesis.

Conditional Generative Adversarial Networks (CGAN) were first proposed by Mirza and Osindero[20] and included a conditional variable \(c\) to both the generator and discriminator, therefore influencing the data generating process by this extra knowledge. StarGAN[6] enabled simultaneous training of many datasets within a single network, therefore addressing the difficulty of picture translation across several domains. Facebook AI expanded Instance-conditioned GAN[5] for class-conditional generation, enabling the control of produced picture semantics by means of the suitable mix of examples and class labels. By defining a target color histogram feature, Afifi et al. [1] managed the colors of GAN-generated pictures using an unsupervised method.

Effective integration of flames into a picture requires a model addressing three main problems. It must first guarantee that flame properties like color and form are both varied and controlled. Second, the produced flame picture ought to be as realistically perceptually as feasible. This implies that the color of the flame, hazy edges, smoke, and reflections on surrounding objects should quite match actual circumstances. Third, the model should move smoothly between the flame and the general scene.

Image-to-image (I2I) translation can be implemented to resolve the aforementioned concerns. Many I2I translation frameworks, including super-resolution[18, 4], semantic synthesis[19, 28], photo enhancement[28, 14], and photo editing [8, 24], have been extensively employed as a result of the recent implementations of GANs[12]. Nevertheless, these frameworks are not appropriate for the purpose of augmenting scenes with blazing flames due to the following reasons. Traditional I2I methods frequently impose a common limitation in which translated images tend to maintain their style and content[19, 28, 22], thereby restricting the variation of shape and color in generated flame images. In addition, the irregularity of latent codes renders image translation unpredictable, despite the fact that certain methods [30] can decompose color and texture features of images into a latent space. Furthermore, methods that depend on explicit codes to regulate image features frequently induce color shifts and background distortions [2]. The resolution of background images is also not maintained by these models, resulting in substantial discrepancies between the flame merging region and the overall scene.

3. Method

3.1 Cycle Generative Adversarial Networks

CycleGAN (Cycle-Consistent Generative Adversarial Network) [29] is designed to learn mapping functions between two distinct domains X and Y. It is particularly useful in scenarios where paired training data (direct correspondences between images in the two domains) is not available. The training samples for CycleGAN consist of two sets: $\{x_i\}_1^N$, where $X_i \in$ X, representing a collection of images from domain X , and$\{y_j\}_{j=1}^M$, where $y_j \in Y$, representing a collection of images from domain Y. The data distributions for the images in the two domains are $x\sim p_{data}(x)$ and $y\sim p_{data}(y)$.As illustrated in Figure 1, CycleGAN consists of two main components: two mapping functions and two adversarial discriminators. The mapping functions are G : X →Y , which learns to map images from domain X to domain Y , and F : Y → X, which learns to map images from domain Y to domain X. The adversarial discriminators are D_X and D_Y. D_X is trained to differentiate between real images from domain X and fake images translated from domain Y by the generator F , ensuring that the translated images{F(y)} look similar to the real images x in domain X . D_Y is trained to differentiate between real images from domain Y and fake images translated from domain X by the generator G , ensuring that the translated images {G(x)} look similar to the real images y in domain Y. CycleGAN employs two primary types of losses to train the model effectively: adversarial losses and cycle consistency losses. Adversarial losses help in aligning the distribution of the generated images with the distribution of the data in the target domain. The adversarial loss for the generator G ensures that the translated images {G(x)}are indistinguishable from the real images y , and similarly, the adversarial loss for the generator F ensures that the translated images {F(y)} are indistinguishable from the real images x. Cycle consistency losses ensure that the mappings G and F do not contradict each other. Specifically, for any image x from domain X, applying G followed by F should ideally bring the image back to the original domain X , i.e., F(G(x)) $\approx x$, and for any image y from domain Y , applying F followed by G should ideally bring the image back to the original domain Y, i.e., G(F(y))$\approx y$. By optimizing these losses, CycleGAN achieves effective and consistent image translation between two domains, even in the absence of paired training data.

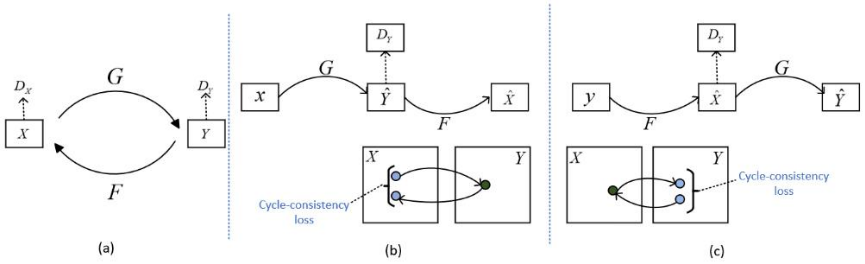

Figure 1: CycleGAN structure；

3.2 Improved Cyclegan Networks

3.2.1 Globally Attachable Residual Networks

Our suggested residual network structure combines self-attention learning with context mining into a single design, hence enhancing the network's capacity to collect contextual information as well as the richness of feature representation. By means of a combination of the localized static context with the self-attention-guided dynamic context, so improving the ability of feature extraction by addressing the defects of the original residual network limited to extracting features in a small window, increases the global field of view of feature extraction, and enhances the network's capacity to understand the global dependencies of the input data.Figures 2 depict the CT_Blocks network architecture.

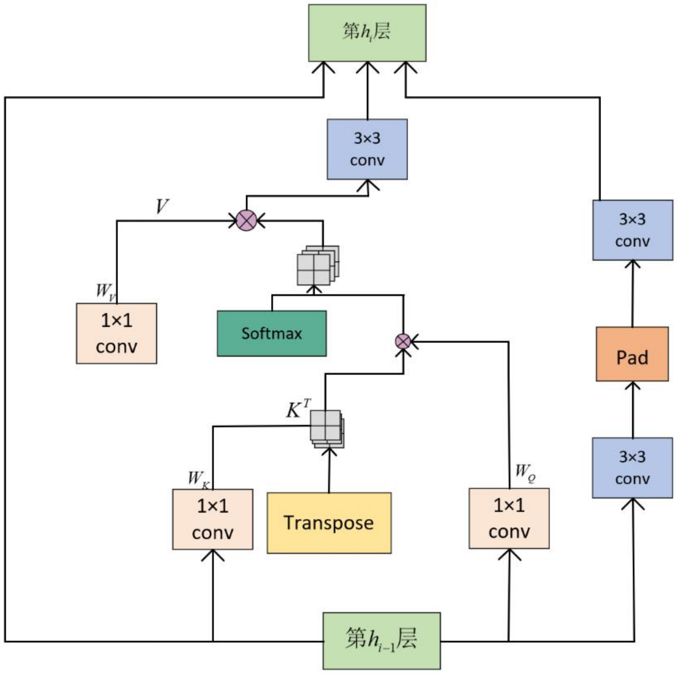

Figure 2 CT_Blocks network structure diagram

The CT_Blocks network structure, in addition to the two 3×3 convolutional layers (Conv) inherited from the original residual network, adds a Contextual Transformer (CoT), which works by taking the feature map X∈[C,W,H] ,C is the number of channels, H is the height and W is the width. The input feature map X is first processed by a 3×3 convolution to encode static context information for each spatial location of the key (key). This produces a static context representation K_1 .The input feature map X is spliced with the static context representation K_1 . to form the query(Q) and key(K).The dynamic multi-head attention matrix is then learned by two consecutive 1$\times$1 convolutions (the first witha ReLU activation function and the second without). Using the learned attention matrix, the values (V) in the input feature map X are weighted and summed to aggregate information from all other locations to obtain the dynamic context representation K_2 , which fuses the static context representation K_1. with the dynamic context representation K_2. This is usually achieved by global average pooling of the channel dimensions and a soft attention mechanism to adaptively aggregate the two kinds of context information. The fused context information is used as the output $X^{'}$ of the CoT block, and this output feature map $X^{'}$ contains rich static and dynamic context information. In order to increase the multi-scale spatial expression capability, $X^{'}$ is again passed through a 3×3 convolutional layer to obtain the output feature map Y, whose expression is:

$$Y = X + Conv\big(CoT(X)\big) + Conv\big(Conv(X)\big)\#(1)$$

3.2.2 generator network structure

The generator model in this study utilizes an Auto-Encoder+Skip-connection network structure. The residual network module deviates from the original residual network structure and instead employs the CT_Blocks structure proposed in this paper. This unique structure provides global connectivity, addressing the limitations of the original residuals that can only extract features locally. As a result, the model is capable of both global feature extraction and multi-scale invariance, leading to improved image quality. The structure of the generator network, as depicted in Figure 3, is presented in this paper.

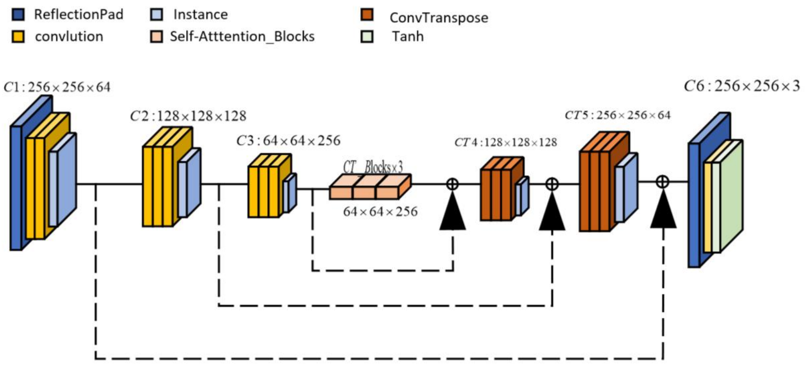

Figure 3 Generator network structure diagram

The dark blue module is the mirror fill layer, which will fill three pixel matrices of mirror content around the image; C_i :W×H×C indicates that it is currently the ith layer of convolution, through which the width of the output feature map is W, the height is H, and the number of channels is C. The yellow module is the convolution operation with convolution kernel size 3 and step size 2. The red module is an inverse convolution operation with convolution kernel size 3 and step size 1/2, which is used to send the convolved feature map larger. The pink module is the 3 consecutive CT_Blocks structure to give the extracted features a global view. The blue module is the IN layer, which normalizes the feature map to prevent overfitting; the dashed part is to fuse the low-level features obtained after convolution with the high-level features at the same resolution in the model, so that the utilization of feature information in each layer of the model can be improved.

In generating fire images, the network structure using Auto-Encoder and Skip-connection can significantly improve the fidelity and diversity of the generated images.Auto-Encoder can accurately capture the detailed features of the fire images, while Skip-connection helps to merge the features at different scales to enhance the model generalization capability.

This structural design not only improves the stability during the training process, but also helps to generate high-quality fire images with complex scenarios and dynamic changes, which is helpful for security-restricted environments that do not allow starting fires to capture fire videos or testing fire detection algorithms. Improving the impact and effectiveness of fire detection algorithms in different environments.

Auto-Encoder, through its encoder and decoder architecture, is able to learn key features in fire images, such as the color, shape, and dynamics of the flames, as well as smoke dispersion patterns. This helps the generator to accurately reproduce the visual characteristics of the fire when creating the image.Skip-connection allows the model to fuse features at different levels, which means that the network can capture the details of the fire at both the macro- and micro-scales, thus generating a richer and more realistic fire scene. Also by creating direct paths in the network, it helps the gradient to propagate more efficiently during the training process, which reduces the problem of gradient vanishing in deep network training, thus improving the training efficiency and stability of the model. Since Auto-Encoder learns a compressed representation of the data, this helps the model capture the generalized features of the fire images, allowing the generated images to be not only limited to specific fire cases, but to adapt to a wide range of different fire scenarios. Multiple representations of fire images can also be learned, which provides a basis for generating diverse fire scenarios. Combined with Skip-connection's multi-scale feature fusion, it is possible to generate a variety of fire images ranging from light smoke to violent flames.

3.2.3 Discriminator Network Structure

The discriminator network, employing an Auto-Encoder network structure instead of the original structure, offers a key advantage: the training of the discriminator is no longer limited by the constraints of the generator. This allows for the initial training of the discriminator, which in turn stimulates the training of the generator through optimization. This addresses the issue of training imbalance in the original model. The architecture of the discriminator network is shown in Figure 4.

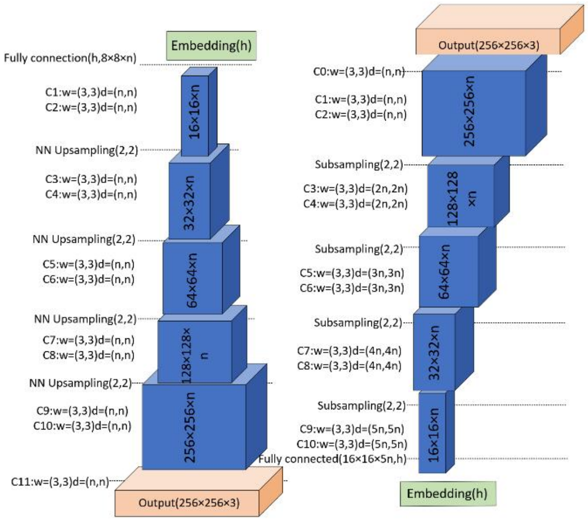

Figure 4 Discriminator network structure diagram

In fire generation, CycleGAN network with Auto-Encoder instead of the original discriminator structure can significantly improve the realism and detail richness of the generated images.The introduction of Auto-Encoder enables the discriminator to be trained independently of the generator, which helps the discriminator to more accurately extract the features of the fire images, such as dynamic changes of the flames and smoke diffusion patterns, thus improving its ability to distinguish between real and generated images. At the same time, training the discriminator independently also improves the stability of the training process, allowing the generator to focus more on producing high-quality fire images. In addition, the Auto-Encoder's multi-scale learning capability helps to capture both micro- and macro-features of the fire, further enhancing the realism of the generated images. This architecture also provides training efficiency gains and flexibility, allowing the model to adapt to different fire scenarios and conditions while addressing possible imbalances in the training process.

3.2.4 Overall Structure

Our goal is to realize the style transformation between fire-free and fire scenes by proposing a new model-SCGAN. The model significantly improves the extraction of fire image features by combining context mining, self-attention learning, and residual network structure. The model employs CT_Blocks, an innovative structure that enhances the global view of features and multi-scale invariance by fusing static contextual information and dynamic multi-head attention matrices. In addition, the generator network utilizes Auto-Encoder and Skip-connection structures, which not only improves the quality of the generated images, but also enhances the generalization ability of the model. The independently trained Auto-Encoder discriminator further improves the accuracy of the discriminator and the stability of the training process. The goal of the whole model is to realize the style transition from fire-free to fire-aware scenarios to support the development and testing of fire detection algorithms, especially in safety-constrained environments where actual ignition is not allowed.

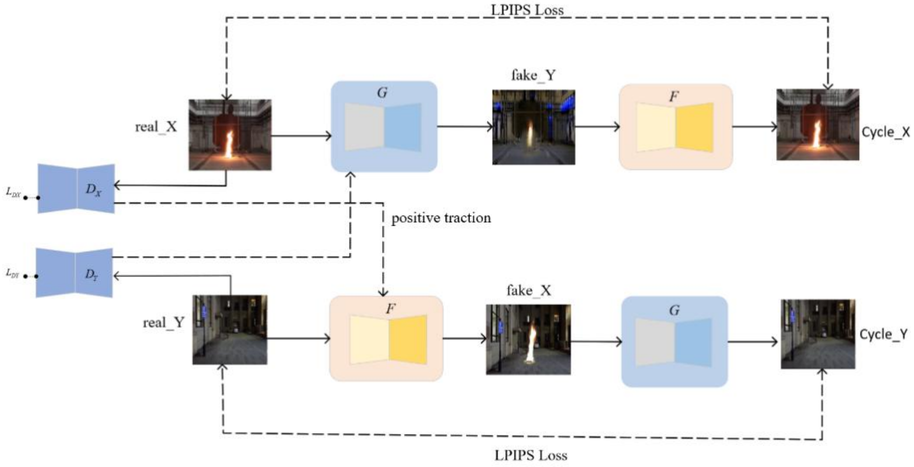

Figure 5 Flow of flame generation and scene migration.

3.3 Denoising Diffusion Probabilistic models

Denoising Diffusion Probabilistic Models (DDPMs) [13, 25] are generative models used for image generation through variational inference. They operate using a Markovian process with a finite number of timesteps 'T'. The process involves two stages: the forward process and the reverse process. In the forward process, a clean image y_0 is sampled, and small Gaussian noises with variance schedules $\{\beta_1, . . . \beta_T, \}$ are added over 'T' timesteps, progressively corrupting the image. In the reverse process, this corrupted image is denoised step-by-step, removing the noise added during the forward process, and ultimately recovering the original clean image.The overall forward process and each forward step are defined as

$$q(y_t|y_{t-1}) = \mathcal{N}\left(y_t; \sqrt{1-\beta_t}y_{t-1}, \sqrt{\beta_t}I\right)\#(2)$$

$$= \sqrt{\beta_t}y_{t-1} + \epsilon\sqrt{1-\beta_t}, \epsilon \sim N(0,I)\#(3)$$

When describing noisy samples, let y_t denote the noisy sample generated at timestep t , and \ y_{t-1} (y_{t-1} \) represent the noisy sample from the previous timestep t-1. The variation in these noisy samples is influenced by the variance schedules β_i, which dictate how the noise evolves over time and thus affect the overall behavior of the samples. Given the Markovian process, where the current state depends only on the previous state and not on the sequence of events leading up to it, the process can be effectively represented by the initial data sample y_0. Specifically, the distribution of the noisy samples results from the sum of t zero-mean Gaussian distributions, each with its own variance schedule. This implies that the evolution of the process can be described using the initial sample y_0 and the noise models.

$$q(y_t|y_0) \coloneqq \mathcal{N}\left(y_t; \sqrt{\bar{\alpha}_t} y_0, \left(1-\bar{\alpha}_t\right) I\right) \#(4)$$

$$= \sqrt{\bar{\alpha}_t} y_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \epsilon \sim \mathcal{N}(0, I) \#(5)$$

where $\alpha_t = 1 - \beta_t$. In the reverse process we define the joint distribution $p_\theta(y_\theta : T)$ with parameters θ. Similar to the forward process, the reverse process is also a Markovian process which is defined as follows

$$p(y_T) \coloneqq N(0, I) \#(6)$$

$$q(y_{t-1}|y_t) = N\left(y_{t-1}; \mu_\theta(y_t, t), \sqrt{\beta}_t I\right) \#(7)$$

To optimize network parameters θ, the objective is to minimize the variational lower bound of the negative log likelihood of the clean image distribution y_0. Training follows the simplified objective for DDPMs proposed in [13]. During training, a timestep t is sampled uniformly from the range T [1, T] , and the noisy sample for this timestep is generated using equation (4), defined by

$$y_t = \sqrt{\bar{\alpha}_t} y_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \epsilon \sim N(0, I) \#(8)$$

The network $P_\theta(.)$ predicts the noise ϵ in this image taking y_t and t as the inputs. The training objective is defined as,

$$L_{simple} \coloneqq E_{t\sim[1,T],\epsilon\sim N(0,I)}[\| \epsilon - \epsilon_\theta(y_t, t)\|^2] \#(9)$$

3.4 Conditional Diffusion Probabilistic Models

The equations previously mentioned are intended to facilitate the generation of images. The conditional distribution of the clean image must be modeled in order to employ DDPMs for low-level vision tasks, such as image restoration. A straightforward method for modeling the conditional distribution of a clean image in relation to its corresponding degraded image has been proposed by Saharia et al. The forward process in conditional DDPM is identical to that of the unconditional model. Random Gaussian noise is introduced to a clear image obtained from the dataset using a randomly sampled timestep t. The degraded image(x) is also provided as input to the neural network during the reverse process, in addition to the chaotic image and the time t. Hence the denoising model is defined by P^θ(y_t, x, t) and the reverse process is defined by

$$p(y_T) = \mathcal{N}(0, I) \quad (10)$$

$$q(y_{t-1}|y_t, x) = \mathcal{N}\left(y_{t-1}; \mu_\theta(y_t, x, t), \sqrt{\beta}_t I\right)\#(11)$$

The mean $\mu_\theta(y_t, x, t)$ is estimated according to,

$$\mu_\theta(y_t, x, t) = \frac{1}{\sqrt{1-\beta_t}}\left(y_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\theta(x, y_t, t)}\right)\#(12)$$

3.5 Proposed method

The method and training process that we suggest are elaborated upon in this section. In order to train our model, we implemented a multi-stage training procedure, as illustrated in Figure 6. Given a dataset containing real flame images, we train a diffusion model, $p_\theta(.)$, to perform the task of generating fire images unconditionally. Our goal is to synthesize more fire images of the scene and generate more realistic fire images with improved DDPM. The image generated by SCGAN may contain some noise and blur, which cannot produce the details and texture of the flame well, and the image detail and clarity can be significantly enhanced through the denoising process of DDPM.

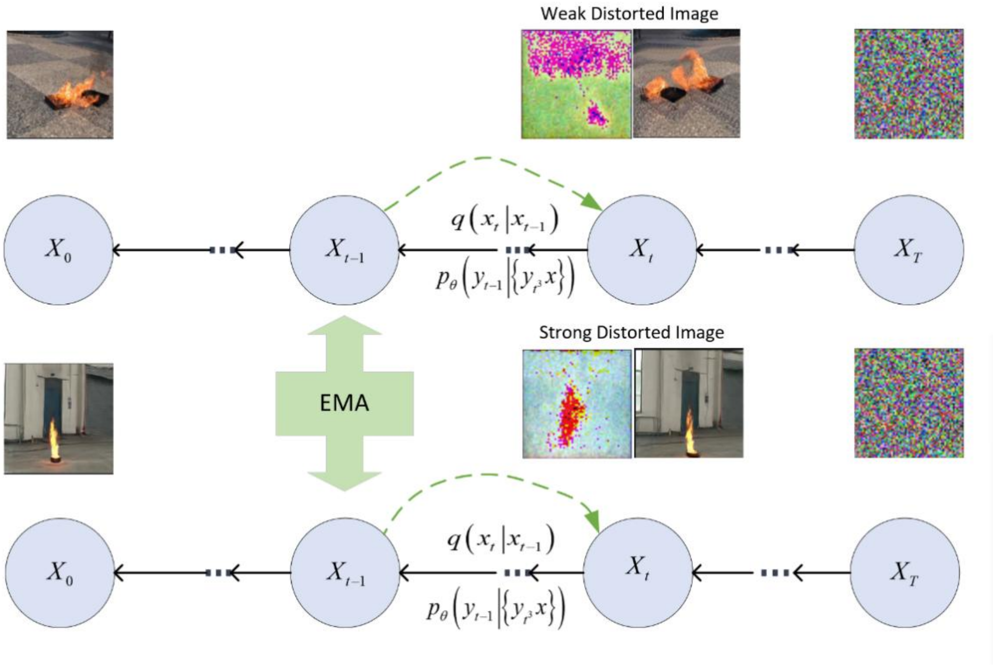

Figure 6 An overview of the proposed approach

To enhance the model's robustness to distorted and blurred images, we utilize a super-resolution approach with two distinct models: $p_\emptyset$ (.)for handling weak distortions and p (.)for recovering from strong distortions. To optimize the parameter δ, we sample a noise vector δ from a Gaussian distribution $\epsilon \sim N(\theta, I)$ and generate a noisy image y_t

from the clean image y_0. The parameter $\delta$ is then optimized in each training iteration using the loss function L_{final}, which evaluates the model's performance in reconstructing images from their distorted versions. This process ensures that the model can effectively recover high-quality images even from significant distortions.

For each training iteration, the parameter $\delta$ is optimized using the loss function L_{final} defined as follows

$$L_T = \parallel \epsilon - \epsilon_\delta(y_t, x_{S-ture}, t) \parallel^2, \epsilon \sim N(0, I), \#(13)$$

$$L_S = \parallel \epsilon_\phi(y_t, x_{W-ture}, t) - \epsilon_\delta(y_t, x_{S-ture}, t) \parallel^2 \#(14)$$

$$L_{final} = E_{t\sim[1,T]}[L_t + \gamma L_S] \#(15)$$

The term L_S ensures the model focuses on accurately reconstructing the flame region despite distortion. Parameter $\delta$ is updated by optimizing the loss function L_{final}. Model weights $\delta$ are updated using an exponential moving average (EMA) rather than directly through the loss function optimization.The EMA-based weight update of the weights $\delta$ using the estimated weights $\emptyset$ is performed according to the

$$\phi = \gamma_1\phi + (1 - \gamma_1)\delta \#(16)$$

In diffusion model inference, conventional methods are time-consuming, but satisfactory results can be achieved with fewer timesteps, typically between T = 40 - 50 . Instead of starting with pure Gaussian noise, inference can begin with a distorted, blurred image, which helps fix rough features and improves efficiency. Let x be the distorted input image and y_t the inference image after t forward steps. The process starts with Gaussian noise at t = T and refines the image from t = t_1 to T . This approach captures flame features effectively and restores details such as edges and texture, resulting in a high-quality image.

4. Experiments

4.1 Implements Details

In our experiments, we use FireNet , the Fismo dataset and the dataset from the self-constructed dataset as the fire dataset. In addition, we sampled and mixed Arnaud Rougetet1's Google Landmark v2 dataset, Landscape Pictures dataset and self-built dataset as clean dataset [27].

Table 1. Details of the datasets for training;

Dataset	Description

FireNet	46 fire videos
Fismo dataset	FireVid,RESCUER Video Dataset
Self-constructed dataset	125 experimental flames and smokevideos

Figure 7 An overview of the proposed approach

4.2 Quantitative Evaluation Metrics

A quantitative evaluation is conducted in this paper on three dimensions: the performance of computer vision algorithms, the similarity between synthesized images and actual photographs, and subjective user feedback. The Fréchet Inception Distance (FID) is employed to assess the similarity between the synthesized and actual images. This metric is obtained by feeding the synthesized images into a pre-trained Inception model and subsequently computing the FID using the features extracted from the penultimate layer. The FID score is indicative of the quality of detail and variety in the synthesized images, as well as the resemblance between the two. A lesser FID score suggests that the image synthesis quality is preferable.

For the computer vision component, the paper employs the ResNet accuracy and the confidence level from the YOLOv8 model. The ResNet accuracy measures how well the fire and non-fire categories are distinguished in the synthesized images by a trained ResNet network, with higher accuracy signifying more convincing image authenticity. The YOLOv8 confidence indicates the likelihood that an object within the synthesized image is identified as fire, as determined by a trained YOLOv8 network. A higher confidence score suggests a more realistic depiction of the image, particularly in the portrayal of fire halos and reflections.

The user evaluation involves a group of 10 users who are tasked with choosing the image with the most effective synthesis from a collection created by various networks. The score for each network is calculated as the proportion of top-performing images synthesized by that network relative to the total number of images synthesized. The evaluation is bifurcated into global and local assessments. Globally, users focus on the overall authenticity of the image composition, while locally, they concentrate on the specifics of fire halo and reflection generation.

4.3 Qualitative Evaluation

The test sets and prediction sets were the subjects of the qualitative evaluation in this paper. The background and flame position of the test sets were identical to those of the training sets, but the flame was distinct. The actual images of the test sets were available. The background, flame, and flame position of the prediction sets were all distinct from those of the training sets, and the actual images were unknown.

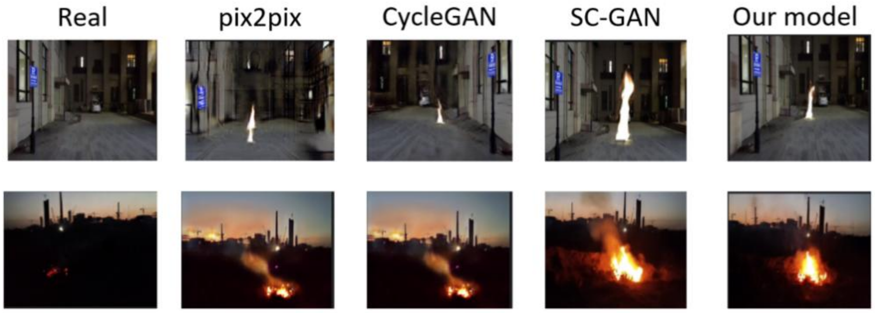

Figure 8 Comparison chart of experimental results

An illustration of the experimental outcomes is illustrated in Figure 8. The CycleGAN synthetic image's reflection regions are arbitrarily larger or smaller than the actual image, as illustrated in Figure 8. The flame halos and reflections in the Pix2pix synthesized image are relatively indistinct, which is a departure from the actual image. The SC-GAN synthesized images exhibit relatively large flames with distinct boundaries between reflections and background, but they differ from the actual images. Our model's flame size and reflection are more reasonable and more consistent with the aesthetics of human vision.

4.4 Quantitative Evaluation

The article conducted comprehensive testing on the images synthesized by different networks, including FID values, ResNet accuracy, YOLOv8 confidence, and user evaluations, with the results summarized in Table 1. In the comparison of FID values, it was found that the two-stage network structure proposed in this paper, with the lowest

FID value of 28.65, outperformed other networks. This indicates that the images synthesized by this structure are superior in terms of the realism, clarity, and diversity of halos and reflections. Further confirmation came from the evaluation of computer vision algorithms, where the two-stage network structure achieved ResNet accuracy and YOLOv8 confidence levels of 0.9568 and 0.7539, respectively, both superior to other networks, demonstrating that the synthesized images are visually capable of "misleading" fire classification and recognition algorithms.

Table2. Quantitative evaluation results;

	CycleGAN	Pix2pix	SC-GAN	Our model
FID	47.1	40.29	48.52	28.46
Computer vision	0.7778	0.7228	0.9105	0.9568
Acc conf	0.6067	0.5788	0.6818	0.7539
User evaluation	0.126	0.095	0.125	0.682
Global local	0.027	0.036	0.179	0.654

In terms of user evaluations, whether considering the overall authenticity of the images or assessing the details of the flame halo and reflections, users generally preferred the images synthesized by the two-stage network structure proposed in this paper. This reflects that the images synthesized by this structure not only align more with human visual aesthetics but are also more likely to "deceive" human visual judgment.

In conclusion, based on all test results, it can be determined that compared to other synthesis methods, the two-stage network structure proposed in this paper has a distinct advantage in the quality of synthesized flame images.

5. Conclusions

The article presents a novel approach to create fire pictures to solve the uneven sample counts between fire and non-fire images in deep learning applications. The approach comprises of two networks: a hybrid network and a flame generating network based on the production of fire properties. By means of form of the outputs depending on any input given as a reference, the flame generating network synthesizes fire pictures of various circumstances and generates a range of fire images using SCGAN technology.

Conversely, by optimizing the texture and style of the hybrid area, the hybrid network synthesizes fire pictures from many situations into an upgraded DDPM to provide realistic visuals.

The method has three main advantages:

1. The flame images generated by controlling with SCGAN can achieve various fire image generations that existing methods cannot, as well as style conversion between fire-free and fire scenes.

2. The flame and background images are further fused by the improved DDPM network, and high-quality background details are preserved in the generated images.

3. Since the model is trained on a real dataset, the generated images can be used in engineering application scenarios for target detection algorithms.

Furthermore, the article also conducted experimental verification of the proposed method. Through quantitative and qualitative assessments, the results show that the two-stage network structure proposed in this paper has a clear advantage in the quality of flame image synthesis. Test results using FID values, ResNet accuracy, YOLOv8 confidence, and user evaluations indicate that the method proposed in this paper is superior to other methods in generating realistic, clear, and diverse flame images.

In summary, the method proposed in this paper not only enhances the expansion capability of the fire image dataset but also helps to improve the effectiveness and adaptability of fire detection systems by generating high-quality fire samples, especially in environments where actual ignition is not allowed to capture fire videos or test fire detection algorithms.

6. References

[1] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7941–7950, 2021.

[2] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. Pirm challenge on perceptual image super-resolution (2018), 2018.

[5] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 8789–8797, 2018.

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[8] Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7902–7911, 2018.

[9] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[10] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] A Ignatov, N Kobyshev, R Timofte, K Vanhoey, and L Wespe Van Gool. Weakly supervised photo enhancer for digital cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 691–700.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[19] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5524–5532, 2018.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[21] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. Recent progress on generative adversarial networks (gans): A survey. *IEEE access*, 7:36322–36333, 2019.

[22] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arxiv 2022. *arXiv preprint arXiv:2204.06125*, 2022.

[24] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017.

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[26] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[30] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017.

[31] Song Chun Zhu, Ying Nian Wu, and David Mumford. Minimax entropy principle and its application to texture modeling. *Neural computation*, 9(8):1627–1660, 1997.